DOCUMENT RESUME

ED 390 931 TM 024 457

AUTHOR Harwell, Michael; Serlin, Ronald
TITLE An Empirical Study of the Type I Error Rates of Five Multivariate Tests for the Single-Factor Repeated Measures Model.
PUB DATE Apr 95
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Chi Square; *Error of Measurement; Monte Carlo Methods; *Multivariate Analysis; *Nonparametric Statistics; *Sample Size; Statistical Distributions
IDENTIFIERS Empirical Research; F Test; Hotellings t; *Rank Order Transformation; Repeated Measures Design; *Type I Errors

ABSTRACT

A Monte Carlo study was used to examine the Type I error rates of five multivariate tests for the single-factor repeated measures model. The performance of Hotelling's T-squared and four nonparametric tests, including a chi-square and an "F" test version of a rank-transform procedure, was investigated for different distributions, sample sizes, and numbers of repeated measures. The results indicated that both Hotelling's T-squared and the F test version of the rank-transform test performed well, producing Type I error rates that were close to the nominal value. The chi-square version of the rank-transform test, on the other hand, performed poorly for virtually all conditions studied. The performance of the other nonparametric tests depended heavily on sample size. Based on these results, Hotelling's T-squared is recommended for the single-factor repeated measures model. Appendix A discusses computing the tests. (Contains 1 table and 48 references.) (Author/SLD)

An Empirical Study of the Type I Error Rates of Five Multivariate Tests for the Single-Factor Repeated Measures Model

Michael Harwell

University of Pittsburgh

Ronald Serlin

University of Wisconsin-Madison

April, 1995

Paper presented at the annual meeting of the American Educational Research Association, April, San Francisco. Correspondence concerning this paper should be directed to Michael Harwell, 5H33 Forbes Quad, University of Pittsburgh, PGH, PA 15260.

Abstract

A Monte Carlo study was used to examine the Type I error rates of five multivariate tests for the single-factor repeated measures model. The performance of Hotelling's T^2 and four nonparametric tests, including a chi-square and an F test version of a rank-transform procedure, was investigated for different distributions, sample sizes, and numbers of repeated measures. The results indicated that both Hotellings T^2 and the F test version of the rank- transform performed well, producing Type I error rates which were close to the nominal value. The chi-square version of the rank-transform test, on the other hand, performed poorly for virtually all conditions studied. The performance of the other nonparametric tests depended heavily on sample size. Based on these results, Hotelling's T^2 is recommended for the single-factor repeated measures model.

An Empirical Study of the Type I Error Rates of Five Multivariate Tests for the Single-Factor Repeated Measures Model

Experimental settings in which i = 1, 2, ..., N subjects (blocks) are measured on P occasions with the same variable are often referred to as repeated measures designs. We consider the unreplicated design in which subjects are treated as a random effect and the repeated factor as a fixed effect. It is well known that both univariate and multivariate normal-theory tests of the (main) effect associated with the repeated factor can be performed, and that both procedures require that the N vectors of errors be independently and (multivariate)-normally distributed (Bock, 1975).

The univariate approach also requires that the covariance matrix of the repeated measures Σ possess sphericity, which exists in the sample if the statistic

$$\varepsilon = \frac{(\mathrm{tr}\mathbf{C}'\Sigma\mathbf{C})^2}{(P-1)\mathrm{tr}(\mathbf{C}'\Sigma\mathbf{C})^2} \qquad (1)$$

equals one; otherwise, the data show some degree of nonsphericity. (The lower bound of ε indicating maximum lack of sphericity is $(P-1)^{-1}$). In equation (1), **C** is a (P-1) x P matrix of coefficients defining a collection of orthonormalized contrasts and **tr** is the trace operator (Box, 1954). If $\varepsilon = 1$, or, in practice, is quite close to 1, the univariate F test is often recommended because of its greater power relative to the multivariate approach (Huynh & Feldt, 1970; Rouanet & Lepine, 1970) However, use of the univariate F when sphericity is violated is known to effect the Type I error rate of F, typically producing inflated error rates (Boik, 1981; Collier, Baker, Mandeville, & Hayes, 1967; Huynh & Feldt, 1980; Mendoza, Toothaker, & Nicewander, 1974). Complicating matters is the problem that repeated measures data can be expected to be nonspherical (Greenwald, 1976; O'Brien & Kaiser, 1985; Romaniuk, Levin, & Hubert, 1977; Wilson, 1975), with ε values often between .75 and .85 (Huynh & Feldt, 1976).

One alternative in the face of nonspherical data is to use an adjusted univariate F test (c.f., Huynh, 1978; Quintanna & Maxwell, 1994; Rogan, Keselman, & Mendoza, 1979); another is to use a multivariate test which makes no assumption about the structure of Σ. Indeed, several authors (e.g., Cole & Grizzle, 1966; Lewis, 1993; Marascuilo & Levin, 1983, p. 381; Maxwell & Delaney, 1990, p. 591; Peng, 1975) have expressed a preference for the multivariate approach, a preference with which we concur. Still, opting for a multivariate test does not settle

things because several such tests are available, including the normal-theory Hotelling's T^2 and various nonparametric tests which do not require normality of the distribution of errors.

Description of the Problem

Akritas and Arnold (1994) provided a theoretical justification for (nonparametric) rank-transform (RT) tests for several designs, including the single-factor repeated measures design. Rank-transform tests rank the raw scores and perform normal-theory tests on the ranks with no assumption that the data are normally distributed. The Akritas and Arnold test uses Hotelling's T^2, computed for ranked data. Interestingly, they made no mention of the fact that their form of T^2 is the same as that proposed by Agresti and Pendergast (1986).

Despite the intuitive appeal and ease of use of RT tests, and the fact that the RT procedure has been embraced in the documentation of the SAS (SAS Inc., 1985, p. 647) statistical analysis program, use of the test proposed by Akritas and Arnold should not go unchallenged for at least three reasons. First, there is some evidence that the Hotelling T^2 is robust under certain conditions, and, thus, can be used with some nonnormal distributions. If competing normal-theory and nonparametric tests show the same statistical behavior for realistic datasets (e.g., nonnormal data), we would opt for the normal-theory test. Second, the validity of RT tests has been questioned, with several papers (e.g., Blair, Sawilosky, & Higgins, 1987; Fligner, 1981; Sawilosky, Blair, & Higgins, 1989) providing evidence of the shortcomings of these tests in certain settings. Third, there are other nonparametric tests which can be used in the single-factor repeated measures design and as such represent important data-analytic alternatives.

In short, before the Akritas and Arnold RT test can be recommended over its competitors there must be evidence supporting its superior statistical properties for realistic data conditions. This paper reports the results of a Monte Carlo study of the Type I error rates (α) of five multivariate tests for the single-factor repeated measures model: Hotelling's T^2, two versions of Akritas and Arnold's RT test, a test due to Puri and Sen (1969), and a multivariate Wilcoxon signed-ranks test (Bickel, 1965; Hettmansperger, 1984, pp. 283-285). Univariate RT tests for the repeated measures model (e.g., Kepner & Robinson, 1988) are not considered.

Data Model and Statistical Tests

Following Davidson (1980), the linear model assumed to underlie the (continuous) data is

$$y_{ip} = \mu + \tau_p + \varepsilon_{ip}, \quad (2)$$

where $\Sigma_p \tau_p = 0$; $E(\varepsilon_{ip})=0$; $cov(\varepsilon_{ip},\varepsilon_{i'p'})= \delta_{ii'} \sigma_{pp'}$ where $\delta_{ii'}=1$ if i=i'and 0 otherwise and $\sigma_{pp'}$ is the covariance. In equation (2), y is the observed score of the ith subject on the Pth repeated measure, μ is a grand mean, τ_p is a treatment effect defined as $\mu_p - \mu$, and ε_{ip} is an error term. We assume that covariances among the errors are collected in the matrix Σ. All of the tests assume that the N vectors of errors are independently distributed.

The hypothesis tested by Hotelling's T^2 is Ho: $\tau_1 = \tau_2 = .. = \tau_p$. The form of the the test statistic is

$$T^2 = N(\mathbf{C}\bar{\mathbf{Y}})'(\mathbf{CSC}')^{-1}(\mathbf{C}\bar{\mathbf{Y}}) \quad (3)$$

For convenience this test is often transformed into an F:

$$F = \frac{(N-P+1)}{(P-1)} \quad \frac{T^2}{(N-1)} \quad (4)$$

Under Ho, the above statistic is distributed as an F with P-1 and N-P+1 degrees of freedom if the N error vectors are multivariate-normally distributed.

The RT procedure of Akritas and Arnold (1994) tests the hypothesis of homogeneity of the marginal distribution functions Ho: $F_1(y)= F_2(y)= ...=F_p(y)$; rejection of this hypothesis implies, but does not guarantee, differences among location parameters. (All of the nonparametric tests in this paper share this null hypothesis). To compute the chi-square version of the Akritas and Arnold test (AACHI) the NP raw scores are ranked from 1 to NP, T^2_{RT} is computed on the ranks, and

$$AACHI = (P-1)\ \frac{T^2_{RT}\,(N-P+1)}{(N-1)} \sim \chi^2_{p-1}\,. \quad (5)$$

Under Ho, the resulting test statistic is asymptotically distributed as a chi-square variable with P-1 degrees of freedom. Agresti and Pendergast (1986) recommended computing AAF = AACHI*(P-1) and comparing this value against an F critical value based on P-1 and N-P+1 degrees of freedom. An advantage of the AACHI and AAF tests is that standard statistical analysis programs can be used to obtain T^2_{RT}; one simply submits the ranks to a program that computes T^2 for repeated measures models. (For all of the nonparametric tests, ties among the raw

scores are handled by assigning midranks, which should not have an adverse effect on these tests unless the proportion of ties is large (Lehmann, 1975, p. 18)).

The general linear model procedure due to Puri and Sen (1969) suggests another nonparametric test in the multivariate repeated measures model. Here P-1 differences are created and ranked from 1 to N(P-1). The hypothesis of homogeneity of the $F_p(y)$ can be tested with

$$PS = (N-1)\theta \sim \chi^2_{p-1} \qquad (6)$$

θ is the eigenvalue obtained from the matrix product of equation (2.26) in Puri and Sen (1969) involving the between-measure cross-products matrix **H** and the total cross-products matrix **T**, and is obtained as the solution to the Pillai-Bartlett eigenvalue problem $\mathbf{HT}^{-1}$. Under the null hypothesis of homogeneity of marginal distribution functions, PS is asymptotically distributed as a chi-square variable with P-1 degrees of freedom.

Another alternative is the multivariate Wilcoxon signed-rank (MWSR) test due to Bickel (1965). Although numerous variations of this procedure have been suggested (e.g., Policello & Hettmansperger, 1976; Utts & Hettmansperger, 1980), we study the traditional form of the MWSR test in which P-1 Wilcoxon signed-rank statistics are computed and a test statistic is formed from this vector and the covariances among the signed-rank statistics. The test statistic is compared to a chi-square variable with P-1 degrees of freedom.

Akritas and Arnold (1994) used data from Johnson and Wichern (1988, p. 219) to illustrate the computations for the AACHI test. We use the same data to illustrate the computations for each of the tests in Appendix A.

Review of the Literature

Surprisingly few studies of multivariate tests for the single-factor repeated measures model have been reported. As might be expected, most of these have investigated Hotelling's T^2.

Jensen (1982) used analytic methods to show that T^2 maintains it Type I error rate for a variety of nonnormal distributions (e.g., t, Cauchy) if some general criteria involving the shape of the distribution are satisfied. Jensen's results help to explain Monte Carlo findings indicating that the Type I error rate of T^2 is robust to symmetric but nonnormal distributions (e.g., Chase & Bulgren, 1971; Serlin & Harwell, 1989; Utts & Hettsmanperger, 1980) and to mild skewing (e.g., Everitt, 1979). Increasingly asymmetric distributions, on the other hand, have sometimes

produced inflated error rates, even as sample size increases (Chase & Bulgren, 1971; Everitt, 1979). For example, Everitt (1979) reported error rates of .14 for $\alpha = .05$ for an exponential distribution, and .30 for a log-normal. Everitt also reported that increasing sample sizes (5, 10, 15, 20) had a limited effect on error rates. Chase and Bulgren's (1971) results for P = 3 showed a similar pattern for T^2 for sample sizes of 5, 10, and 20. However, Serlin and Harwell (1989) found that T^2 was robust for exponential data and a sample size of 30, and concluded "Unlike many simulation experiments, the Type I error results were quite unambiguous, and, for the conditions of this study, provide a textbook example of a robust test." (p. 13) This discrepancy among studies of the robustness of T^2 distributions persists for both equal and unequal between-measure correlations.

Few Monte Carlo studies of nonparametric tests for the repeated measures model have been reported. Agresti and Pendergast (1986) found that the AAF test maintained its Type I error rate for a multivariate-normal distribution for sample sizes of 10, 30 and 50 and P = 2 versus 5 repeated measures. Serlin and Harwell (1989) reported similar findings for the AAF test for N = 30, 100 and P = 3, 4 for a normal, double-exponential, and exponential distributions. Serlin and Harwell also reported that the error rates of the PS test under these conditions were quite conservative.

Design of the Monte Carlo Study

Ideally, the Type I error behavior of the various tests would be investigated analytically. However, such solutions are difficult because they almost always require multivariate-normality, the very assumption that empirical data can be expected to violate. In addition, the nonparametric procedures are large sample tests, and their behavior for small samples must be investigated empirically. We settled for a Monte Carlo study comparing the Type I error rates of the five tests.

Hoaglin and Andrews (1975), Lewis and Orav (1989), and others have argued that Monte Carlo studies should be subject to the same principles of experimental design and data analysis as empirical studies. Accordingly, the design of our simulation study was an unreplicated 5 (type of distribution) x 3 (sample size) x 2 (number of repeated measures) fixed effects, fully-crossed factorial. Type of distribution, sample size, and number of repeated measures served as independent variables and the empirical Type I error rates as the

dependent variable. This design made it possible to examine the empirical error rates for evidence of interactions among the simulation factors and to estimate the magnitude of significant effects.

The simulation factors and factor levels were selected because of their known (or suspected) effects on the Type I error rates of one or more of the tests, and because these factors have been used in previous Monte Carlo studies of the repeated measures model. Table 1 outlines the factors and their levels which were manipulated. The focus on the effects of increasing asymmetry arose from the effect of this factor in previous Monte Carlo studies of the T^2 test. The γ_1 (skewness) = γ_2 (kurtosis) = 0 case produced normally distributed data which acted as a baseline against which other results could be compared, whereas increments of .5 for γ_1 permitted the detection of trends in the empirical error rates for increasingly skewed data (γ_2 was not a focus of the simulation study because there is little evidence that it affects tests of location). A $\gamma_1 = .5$, $\gamma_2 = 1.5$ pairing produced a mildly skewed and somewhat leptokurtic distribution, $\gamma_1 = 1$, $\gamma_2 = 3$ a moderately skewed and leptokurtic distribution which is equal to a chi-square with $v = 8$ degrees of freedom, $\gamma_1 = 1.5$, $\gamma_2 = 4.5$ a skewed and leptokurtic distribution, and $\gamma_1 = 2$, $\gamma_2 = 6$ a badly skewed and peaked distribution which is equal to a chi-square with $v = 2$, or, equivalently, an exponential distribution. The chosen sample sizes of 9, 15, and 30 were intended to reflect quite small to moderate sample sizes that have been used in previous Monte Carlo studies of this model (Chase & Bulgren, 1971; Everitt, 1979; Serlin & Harwell, 1989). The same resoning led to the selection of the P = 3, 4 numbers of repeated measures.

Data Generation

A Gateway DX2/50 microcomputer was used to generate data. All programming was done in FORTRAN IV and was supplemented by subroutines written by the second author. The random number generator was taken from **Numerical Recipes** (Press, Flannery, Teukolsky, & Vetterling, 1986), with model (1) serving as the underlying data generation model. The following steps were followed to generate data: (a) NP scores representing multivariate-normal data were simulated using the Kaiser and Dickman (1962) procedure and, when appropriate, were transformed to nonnormal data using the method of Vale and Maurelli (1983). Habib and Harwell (1989) provide details on using the Vale and Maurelli procedure, which combines the Kaiser and Dickman approach with

Fleishman's (1978) procedure for generating nonnormal data through skewness and kurtosis parameters. In all cases, the between-measure correlations equaled .5 and τ_j equaled 0. (b) Step (a) was repeated 10,000 times and for each replication the T^2, AACHI, AAF, PS and MWSR tests were computed and the test statistics compared to the appropriate critical value for the .05 and .01 levels of significance.

Results

Adequacy of the Data Generation

The adequacy of the data generation was judged by examining the average skewness, kurtosis, and correlation values computed for the simulated data for each combination of conditions, as well as across all conditions. Results for the N = 9, P = 3 case for the various distributions are reported in Table 2, along with overall summary statistics. We report the N = 9, P = 3 case because problems in producing data with the desired properties are likely to be most acute for smaller sample sizes. The results in Table 2 suggest that the simulated data possessed (approximately) the desired marginal skewness, kurtosis, and correlation values. A similar pattern was observed for the larger sample size conditions.

Analysis of the Empirical Type I Error Rates

The empirical Type I error rates are reported in Table 3. Because of the similarity of the results for the .01 and .05 levels, only the latter are reported. The expression $.05 \pm 1.96[(.05(1-.05))/10,000]^{1/2}$ was used to establish a sampling error range for the empirical proportions of rejections. Error rates exceeding the upper limit of .054 were considered to be inflated and are indicated in Table 3 by a *, and error rates below the lower limit of .046 were considered to be conservative and are indicated by a **.

The results in Table 3 suggests the following conclusions: (a) Hotelling's T^2 and the AAF test did the best job of controlling Type I error rates near the nominal value, (b) The AACHI test performed particularly poorly, (c) The PS test was extremely conservative and the MWSR test somewhat less so for larger samples.

It is possible that simple descriptive analyses of the empirical error rates may conceal important information such as the presence of interactions among simulation factors with respect to the empirical Type I errors.

Accordingly, the error rates were analyzed for each test using an unreplicated, three-factor, completely between-subjects ANOVA. The three-way interaction variation was used as an estimate of error. There is some evidence that using the highest-order interaction term in this fashion in the analysis of Monte Carlo results has little effect on the results (Alaysin, 1991). Those results which were significant at the .05 level and whose estimate of effect size exceeded .10 are reported in Table 4. Effect sizes were estimated using Fisher's correlation ratio (sum of squares for that effect divided by the sum of squares total) and the ω^2 statistic (Hays, 1973, p. 485). Because the η^2 and ω^2 indices did not differ by more than .02 on any effect, only η^2 is reported in Table 4. The two effects whose η^2 was $\leq .10$ (.03 and .07) were deemed too small to pursue further.

Interestingly, all of the significant effects reported in Table 4 are main effects, and all produced at least moderate and occasionally quite large η^2 values. Only Hotelling's T^2 was sensitive to type of distribution, a result which is consistent with the Monte Carlo results of Chase and Bulgren (1971) and Everitt (1979); however, the marginal mean error rates for T^2 of $\bar{Y}_{\gamma1=\gamma2=0} = .049$, $\bar{Y}_{\gamma1=.5,\gamma2=1.5} = .048$, $\bar{Y}_{\gamma1=1,\gamma2=3} = .046$, $\bar{Y}_{\gamma1=1.5,\gamma2=4.5} = .047$, and $\bar{Y}_{\gamma1=2,\gamma2=6} = .041$ suggests that η^2 depended heavily on error rates associated with an exponential distribution. In fact, the distribution effect is not significant if error rates for the exponential distribution are deleted. The Type I error rates of the AACHI, PS, and MWSR tests proved to be sensitive to sample size, producing marginal means of $\bar{Y}_{N=9} = .127$, $\bar{Y}_{N=15} = .092$, and $\bar{Y}_{N=30} = .069$ for the AACHI test, $\bar{Y}_{N=9} = .001$, $\bar{Y}_{N=15} = .006$, and $\bar{Y}_{N=30} = .012$ for the PS test, and $\bar{Y}_{N=9} = .019$, $\bar{Y}_{N=15} = .035$, and $\bar{Y}_{N=30} = .044$ for the MWSR test. Similarly, the error rates of the AAF and MWSR tests proved to be sensitive to P, producing marginal means of $\bar{Y}_{P=3} = .051$ and $\bar{Y}_{P=4} = .047$ for the AAF test and $\bar{Y}_{P=3} = .040$ and $\bar{Y}_{P=4} = .026$ for the MWSR test.

Conclusions

The results of this study suggest that, for the conditions studied, researchers concerned with controlling Type I error rates can use either Hotelling's T^2 or the F test version of the Akritas and Arnold (1994) rank-transform statistic in testing for a main effect in the single-factor repeated measures model. Although the F test version of the rank-transform statistic performed well, our preference is for Hotelling's T^2 test because of its use of raw scores as opposed to ranks and because of its membership in the general linear model family of statistical procedures.

The performances of the multivariate Wilcoxon signed-ranks test and the Puri and Sen test were far less impressive. Both of these tests produced quite conservative error rates for smaller sample sizes (especially the Puri and Sen test) which, other things being equal, would be expected to be associated with depressed power values. The chi-square statistic presented in Akritas and Arnold (1994) performed poorly for all conditions and is not recommended.

References

Alaysin, M. (1991). **An empirical investigation of the behavior of some parametric and nonparametric tests for frequently encountered data in educational research.** Unpublished doctoral dissertation, University of Pittsburgh.

Agresti, A., & Pendergast, J. (1986). Comparing mean ranks for repeated measures data. **Communications in Statistics-Theory and Methods, 15,** 1417-1433.

Akritas, M.G., & Arnold, S.F. (1994). Fully nonparametric hypotheses for factorial designs I: Multivariate repeated measures designs. **Journal of the American Statistical Association, 89,** 336-343.

Bickel, P.J. (1965). On some asymptotically nonparametric competitors of Hotelling's T^2. **Annals of Mathematical Statistics, 36,** 160-173.

Blair, R.C. Sawilosky, S.S., & Higgins, J.J. (1987). Limitations of the rank transform statistics. **Communications in Statistics-Simulation and Computation, 16,** 1133-1145.

Bock, R.D. (1975). **Multivariate statistical methods in behavioral research.** New York: McGraw-Hill.

Boik, R.J. (1981). A priori tests in repeated measures designs: Effects of nonsphericity. **Psychometrika, 46,** 241-255.

Box, G.E.P. (1954). Some theorems on quadratic formws applied in the study of analysis of variance problems, II. Effects of inequality of variance and correlation between errors in the two-way classification. **Annals of Mathematical Statistics, 25,** 484-498.

Chase, G.R., & Bulgren, W.G. (1971). A Monte Carlo investigation of the robustness of T^2. **Journal of the American Statistical Association, 66,** 499-502.

Cole, J.W.L., & Grizzle, J.E. (1966). Applications of multivariate analysis to repeated measurements experiments. **Biometrics, 22,** 810-828.

Collier, R.O.J., Baker, F.B., Mandeville, G.K., & Hayes, T.F. (1967). Estimates of test size for several test procedures based on conventional variance ratios in the repeated measures design. **Psychometrika, 32,** 339-353.

Davidson, M. L. (1980). The multivariate approach to repeated measures. BMDP Technical Report #75.

Everitt, B.S. (1979). A Monte Carlo investigation of the robustness of Hotelling's one- and two-sample T^2 Tests. **Journal of the American Statistical Association, 74,** 48-51.

Fleishman, A. (1978). A method for simulating nonormal distributions. **Psychometrika, 43,** 521-532.

Fligner, M.A. (1981). Comment. **The American Statistician, 35,** 131-132.

Greenwald, A.G. (1976). Within-subjects designs: To use or not to use? **Psychological Bulletin, 83,** 314-320.

Habib, A.R., & Harwell, M.R. (1989). An empirical study of the Type I error rate and power for some selected normal-theory and nonparametric tests of the independence of two sets of variables. **Communications in Statistics-Simulation and Computation, 18,** 793-826.

Hays, W.L. (1973). **Statistics for the social sciences**. New York: Holt, Rinehart, and Winston.

Hettmansperger, T.P. (1984). **Statistical inference based on ranks**. New York: Wiley.

Hoaglin, D.C., & Andrews, D.F. (1975). The reporting of computation-based results in statistics. **The American Statistician, 29**, 122-126.

Huynh, H. (1978). Some approximate tests for repeated measurement designs. **Psychometrika, 43**, 161-175.

Huynh, H., & Feldt, L.S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact F-distributions. **Journal of the American Statistical Association, 65**, 1582-1589.

Huynh, H., & Feldt, L.S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. **Journal of Educational Statistics, 1**, 69-92.

Huynh, H., & Feldt, L.S. (1980). Performance of traditional F tests in repeated measures designs under covariance heterogeneity. **Communications in Statistics-Theory and Methods, 1**, 61-74.

Johnson, R.A., & Wichern, D.W. (1988). **Applied multivariate analysis** (2nd. ed.). Englewood cliffs, NJ: Prentice-Hall.

Kaiser, H.F., & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. **Psychometrika, 27**, 179-182.

Kepner, J.L., & Robinson, D.H. (1988). Nonparametric methods for detecting treatment effects in repeated-measures designs. **Journal of the American Statistical Association, 83**, 456-461.

Lehmann, E.L. (1975). **Nonparametrics: Statistical methods based on ranks**. San Francisco: Holden-Day.

Lewis, C. (1993). Analyzing means from repeated measures data. In G. Keren and C. Lewis (Eds.), **A handbook for data analysis in the behavioral sciences**. Hillsdale, NJ: Lawrence Erlbaum.

Lewis, P.A.W., & Orav, E.J. (1989). **Simulation methodology for statisticians, operations analysts, and engineers**. (Vol. I). Pacific Grove, CA: Wadsworth and Brooks/Cole.

Marascuilo, L.A., & Levin, J.R. (1983). **Multivariate statistics in the social sciences: A researcher's guide**. Monterey, CA: Brooks/Cole.

Maxwell, S.E., & Delaney, H.D. (1990). **Designing experiments and analyzing data: A model comparison perspective**. Belmont, CA: Wadsworth.

Mendoza, J.L., Toothaker, L.E., & Nicewander, W.A. (1974). A monte carlo comparison of the univariate and multivariate methods for the groups by trials repeated measures design. **Multivariate Behavioral Research, 9**, 165-178.

O'Brien, R.G., & Kaiser, M.K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. **Psychological Bulletin, 97**, 316-333.

Peng, S.S. (1975, April). Analysis of repeated-measures data. Manuscript for an AERA Mini-Training session, Washington, D.C.

Policello, G.E., & Hettmansperger, T.P. (1976). Adaptive robust procedures for the one sample location problem. **Journal of the American Statistical Association, 71**, 624-633.

Press, W.H., Flannery, B.P., Teukolsky, S.A., & Vetterling, W.T. **Numerical Recipes**. Boston, MA: Cambridge University Press.

Puri, M.L., & Sen, P.K. (1969). A class of rank order tests for a general linear hypothesis. **Annals of Mathematical Statistics, 40**, 1325-1343.

Quintana, S. M., & Maxwell, S.E. (1994). A Monte Carlo comparison of seven ε-adjustment procedures in repeated measures designs with small sample sizes. **Journal of Educational Statistics, 19**, 57-71.

Rogan, J.C., Keselman, H.J., & Mendoza, J.L. (1979). Analysis of repeated measurements. **British Journal of Mathematical and Statistical Psychology, 32**, 269-286.

Romaniuk, J.G., Levin, J.R., & Hubert, L.J. (1977). Hypothesis-testing in repeated-measures designs: On the road map not taken. **Child Development, 48**, 1757-1760.

Rouanet, H., & Lepine, D. (1970). Comparisons between treatment in repeated-measures designs: ANOVA and multivariate methods. **British Journal of Mathematical and Statistical Psychology, 23**, 147-163.

SAS Institute, Inc. (1985). **SAS user's guide: Statistics**. Cary, NC: SAS Institute.

Sawilosky, S.S., Blair, R.C., & Higgins, J.J. (1989). An investigation of the Type I error and power properties of the rank transform proceure in factorial ANOVA. **Journal of Educational Statistics, 14**, 255-267.

Serlin, R.C., & Harwell, M.R. (1989, April). A comparison of Hotelling's T^2 and Puri and Sen's rank test for the single-factor, repeated measures design. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Utts, J.M., & Hettmansperger, T.P. (1980). A robust class of tests and estimates for multivariate location. **Journal of the American Statistical Association, 75**, 939-946.

Wilson, R.S. (1975). Analysis of developmental data: Comparison among alternative methods. **Developmental Psychology, 11**, 676-680.

Vale, C.D., & Maurelli, V.A. (1983). Simulating multivariate nonnormal distributions. **Psychometrika, 48**, 465-471.

Table 1*
Outline of the Simulation Study

	Independent Variables	
Type of Distribution	N	P
Normal (γ_1=0, γ_2=0)	9	3,4
	15	3,4
	30	3,4
Slightly skewed and leptokurtic (γ_1=.5, γ_2=1.5)	9	3,4
	15	3,4
	30	3,4
Moderately skewed and leptokurtic (γ_1=1, γ_2=3)	9	3,4
	15	3,4
	30	3,4
Skewed and leptokurtic (γ_1=1.5, γ_2=4.5)	9	3,4
	15	3,4
	30	3,4
Strongly skewed and leptokurtic (γ_1=2, γ_2=6)	9	3,4
	15	3,4
	30	3,4

+Note. γ_1 = skewness, γ_2 = kurtosis, N = sample size, P = number of repeated measures.

Appendix A
Computing the Tests

Johnson & Wichern (1988, p. 219) used a dataset involving measurements of time (in milliseconds) between heartbeats, which was measured 4 times for 19 dogs. The raw data were:

	Repeated Measures			
Dog	1	2	3	4
1	426	609	556	600
2	253	236	392	395
3	359	433	349	357
4	432	431	522	600
5	405	426	513	513
6	324	438	507	539
7	310	312	410	456
8	326	326	350	504
9	375	447	547	548
10	286	286	403	422
11	349	382	473	497
12	429	410	488	547
13	348	377	447	514
14	412	473	472	446
15	347	326	455	468
16	434	458	637	524
17	364	367	432	469
18	420	395	508	531
19	397	556	645	625

Hotelling's T^2 Test

$$\bar{Y} = \begin{matrix} 368.21 \\ 404.63 \\ 479.26 \\ 502.89 \end{matrix} \qquad S = \begin{matrix} 2819.29 \\ 3568.42 & 7963.14 \\ 2943.49 & 5303.98 & 6851.32 \\ 927.62 & 914.54 & 7557.44 \end{matrix}$$

$\bar{Y}$ is a P x 1 vector of sample means and **S** a P x P covariance matrix. The hypothesis to be tested is Ho: $\tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$, with N = 19 and P = 4. Johnson and Wichern transformed the P repeated measures into P-1 new variables that contained all the between-measure information in the original variables. Any number of transformations will do; we follow Johnson and Wichern and use:

$$C = \begin{matrix} -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{matrix}$$

The sample means are transformed directly:

$$C\bar{Y} = \begin{matrix} 209.31 \\ -60.05 \\ -12.79 \end{matrix}$$

For the covariance matrix S,

$$\mathbf{CSC}' = \begin{matrix} 9432.32 & & \\ 1098.92 & 5195.84 & \\ 927.62 & 914.54 & 7557.44 \end{matrix}$$

Then

$T^2 = N(C\bar{Y})'(CSC')^{-1}(C\bar{Y}) = 19(6.11) = 116.$

As an F,

$$F = \frac{(N-P+1)}{(P-1)(N-1)} T^2 = (.2963)(116) = 34.37.$$

Akritas and Arnold Rank-Transform Test (AACHI)

Ranked data

34.5	73	69.5	71.5
2	1	23	24.5
17	40	13.5	16
38.5	37	62	71.5
28	34.5	59.5	59.5
7	42	57	65
5	6	29.5	47
9	9	15	56
20	44.5	66.5	68
3.5	3.5	27	33
13.5	22	52.5	55
36	29.5	54	66.5
12	21	44.5	61
31	52.5	51	43
11	9	46	49
41	48	75	63
18	19	38.5	50
32	24.5	58	64
26	69.5	76	74

The vector of rank means and the covariance matrix of the rank variables are:

$$\bar{\mathbf{R}} = \begin{matrix} 20.26 \\ 30.82 \\ 48.32 \\ 54.61 \end{matrix} \qquad \mathbf{S}_{rank} = \begin{matrix} 162.253 & & & \\ 182.022 & 447.323 & & \\ 172.476 & 292.628 & 371.333 & \\ 113.875 & 172.472 & 253.796 & 261.792 \end{matrix}$$

First compute

$T^2_{RT} = N(C\bar{R})'(CS_{rank}C')^{-1}(C\bar{R}) = 119.323,$

AACHI = $\{T^2_{RT}\ (N-P+1)/(N-1)\}/(P-1) = \{(119.323)(.0889)\}/(3) = 35.35.$

As an F,

AAF = AACHI*(P-1) = 35.35*3 = 106.05

Puri and Sen Test (PS)

Create P-1 difference variables via the transformation **CY'**, where **Y** is an N x P matrix of the raw scores. The resulting difference variable scores are:

	Difference Variables		
Subject	d_1	d_2	d_3
1	121 (40)	-227 (1)	-139 (5.5)
2	298 (56)	14 (28)	20 (30)
3	-86 (10)	-82 (11.5)	-66 (16)
4	259 (52)	-77 (13)	79 (38)
5	195 (43)	-21 (22.5)	-21 (22.5)
6	284 (55)	-146 (4)	-82 (11.5)
7	244 (49)	-48 (18)	44 (35)
8	202 (45)	-154 (3)	154 (41)
9	273 (54)	-73 (14)	-71 (15)
10	253 (51)	-19 (24)	19 (29)
11	239 (48)	-57 (17)	-9 (25)
12	196 (44)	-40 (19.5)	78 (37)
13	236 (47)	-96 (8)	38 (34)
14	33 (31)	-35 (21)	-87 (9)
15	250 (50)	8 (27)	34 (32.5)
16	269 (53)	89 (39)	-137 (7)
17	170 (42)	-40 (19.5)	34 (32.5)
18	224 (46)	2 (26)	48 (36)
19	317 (57)	-139 (5.5)	-179 (2)

The value in parentheses are ranks. The vector of rank means is

$$\bar{R}_d = \begin{matrix} 45.95 \\ 16.92 \\ 24.13 \end{matrix} \qquad \Sigma_d = \begin{matrix} 116.208 & & \\ 14.28 & 99.202 & \\ 14.149 & 17.472 & 159.264 \end{matrix}$$

Solving the Pillai-Bartlett eigenvalue problem produces $\theta = .96$, so

PS = (N-1)θ = (18)(.96) = 17.23

Multivariate Signed-Ranks Wilcoxon Test (MWSR)

The test statistic is MWSR = $\mathbf{T}'(\mathbf{V})^{-1}\mathbf{T}$, where **T** is a P-1 vector of Wilcoxon signed-rank statistics divided by (N+1) and **V** is the covariance matrix among these statistics. First the univariate Wilcoxon signed-rank statistic is computed for each of the P-1 difference variables, divided by (N+1), and stored in **T**:

$$\mathbf{T} = \begin{matrix} 9.4 \\ 1 \\ 4.15 \end{matrix}$$

To compute **V** we first compute the main diagonal elements, which are simply $v_{t,t} = N(2N+1)/(6(N+1)) = 6.175$. The covariances are computed by adding the cross-product of the signed-ranks and dividing by $(N+1)^2$. Here

$$\mathbf{V} = \begin{matrix} 6.175 & -.354 & -.323 \\ -.354 & 6.175 & .618 \\ -.323 & .618 & 6.175 \end{matrix}$$

Then MWSR = 18.